Metabarcoding and Metagenomics 4: 1–18
DOI 10.3897/mbmg.4.48281

Research Article

Estimation of the relative abundance of species in artificial mixtures of insects using low-coverage shotgun metagenomics

Lidia Garrido-Sanz[1], **Miquel Àngel Senar**[1], **Josep Piñol**[1,2]

1 *Univ. Autònoma Barcelona, Cerdanyola del Vallès, 08193, Spain*

2 *CREAF, Cerdanyola del Vallès, 08193, Spain*

Corresponding author: *Lidia Garrido-Sanz* (Lidia.Garrido@uab.es)

Academic editor: *Xin Zhou* | **Received** 7 November 2019 | **Accepted** 24 December 2019 | **Published** 7 February 2020

Abstract

Amplicon metabarcoding is an established technique to analyse the taxonomic composition of communities of organisms using high-throughput DNA sequencing, but there are doubts about its ability to quantify the relative proportions of the species, as opposed to the species list. Here, we bypass the enrichment step and avoid the PCR-bias, by directly sequencing the extracted DNA using shotgun metagenomics. This approach is common practice in prokaryotes, but not in eukaryotes, because of the low number of sequenced genomes of eukaryotic species. We tested the metagenomics approach using insect species whose genome is already sequenced and assembled to an advanced degree. We shotgun-sequenced, at low-coverage, 18 species of insects in 22 single-species and 6 mixed-species libraries and mapped the reads against 110 reference genomes of insects. We used the single-species libraries to calibrate the process of assignation of reads to species and the libraries created from species mixtures to evaluate the ability of the method to quantify the relative species abundance. Our results showed that the shotgun metagenomic method is easily able to set apart closely-related insect species, like four species of *Drosophila* included in the artificial libraries. However, to avoid the counting of rare misclassified reads in samples, it was necessary to use a rather stringent detection limit of 0.001, so species with a lower relative abundance are ignored. We also identified that approximately half the raw reads were informative for taxonomic purposes. Finally, using the mixed-species libraries, we showed that it was feasible to quantify with confidence the relative abundance of individual species in the mixtures.

Key Words

Eukaryotes, Metazoa, genome skimming, PCR-free, mock sample, sequenced genomes

Introduction

Metabarcoding is a technique used to quantify species abundance in natural communities using high-throughput DNA sequencing. There is usually a PCR step to enrich the DNA of a certain genomic region before sequencing, in what is normally termed amplicon metabarcoding. This technique is well-established and used in many ecological settings and with different groups of organisms (Clare et al. 2016; Evans et al. 2016; Deagle et al. 2018; Taberlet et al. 2018), but there is ample evidence today that this method is not always quantitative (Lamb et al. 2018; Piñol et al. 2019).

There is some consensus in literature that, if the PCR step could be avoided, then the metabarcoding process would be much more quantitative (Taberlet et al. 2012; Zhou et al. 2013; Bista et al. 2018). One PCR-free approach is shotgun metagenomics, where the extracted DNA is sequenced directly, so all PCR-generated biases are avoided (Taberlet et al. 2012; Yu et al. 2012; Zhou et al. 2013; Elbrecht and Leese 2015). This method provides reads from every part of the genome that can be compared with sequences stored in genomic repositories. In prokaryotes, shotgun metagenomics provides more accurate taxonomic identification than the classical 16S amplicon metabarcoding (Chen and Pachter 2005). How-

Copyright *Lidia Garrido-Sanz et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ever, in eukaryotes, shotgun metagenomics is hindered by the scarcity of eukaryote species with sequenced genomes (8,417 eukaryote versus 203,148 prokaryote genomes; NCBI database, accessed on 29 April 2019) and their larger size (400.2 Mb ± 1,106.2 Mb in eukaryotes and 3.9 Mb ± 3.7 Mb on prokaryotes).

In eukaryotes, shotgun metagenomics has been mainly applied using chloroplasts and mitochondrial genomes (Srivathsan et al. 2015; Tang et al. 2014), but also with multiple-copies nuclear genomic regions (Linard et al. 2015). The studies, using mitochondrial genomes, showed that quantitative information could be obtained from heterogeneous samples (Zhou et al. 2013; Tang et al. 2015; Bista et al. 2018), but it is fair to assume that the use of complete genomes would provide better quantitative results. Still, the whole-genome metagenomic approach, albeit conceptually sound, has never been tested in eukaryotes. There are good reasons for the lack of such studies, the most important one being the low number of species with assembled genomes. However, the number of sequenced genomes is quickly increasing, as there are several ongoing projects devoted to obtain complete genomes of several groups of organisms: G10K for vertebrates (Genome 10K community of scientists 2009), GIGA for marine invertebrates (GIGA Community of Scientists 2014), GAGA for ants (Boomsma et al. 2017), i5K for arthropods (Robinson et al. 2011; Levine 2011; i5K Consortium 2013), 10KP for plants (Cheng et al. 2018) and 1KFG for fungi (Grigoriev et al. 2014), amongst others. There is even a proposal to sequence the genomes of all eukaryotic species in ten years for ca. 3 billion dollars (Lewin et al. 2018); this estimate could be optimistic, but it probably means that the objective is within reach in a few decades, not more.

In this paper, we imagine a world in which the complete genomes of all the species are known. We simulate this future world by preparing a reference database of insect species whose genome is already assembled to an advanced degree and available at the NCBI RefSeq repository. We shotgun-sequenced DNA from some of these species in low-coverage single-species libraries, prepared without any PCR step and develop the bioinformatic algorithms necessary to go from raw reads to species assignation. Subsequently, we apply our approach to known mixtures of insect DNA to see if the method produces a quantitative estimate of the insect species present.

This exercise is a preliminary test of the difficulties likely to be faced in the future when an important number of complete genomes becomes available. In particular, we address here the following questions: (1) Is the metagenomic method useful to set apart closely related insect species; (2) What is the proportion of reads that is truly informative for species identification; (3) How many reads are necessary to achieve a reasonable level of confidence to provide quantitative estimates of the relative species abundance?

Material and methods

Reference genomes

We considered all insect species whose genome was sequenced, assembled and available at the NCBI RefSeq Database on 2 August 2018. In total, 115 representative genomes of insect species were downloaded; of those genomes, five were removed for different reasons (Suppl. material 2). The remaining 110 species belonged to 7 orders and 43 families; 28 of them were of the genus *Drosophila*.

Selection of the species, preparation of the DNA libraries and sequencing

From this group of 110 species, we selected 18 species for low-coverage genome sequencing (Table 1), based on availability of fresh specimens. In general, the specimens were captured alive, but for two dipterans, *Ceratitis capitata* and *Bactrocera oleae*, that came together from fly traps and for the bed bug *Cimex lectularius* that was captured and stored by a pest-control company. The specimens were preserved in 70% ethanol at 4ºC for no longer than a few weeks and high quality DNA was extracted from ca. 20 mg of material of each species. In some cases, multiple extractions were done to obtain the minimum amount of DNA required for library preparation. We used the DNeasy Blood & Tissue Kit (Qiagen) to extract the DNA.

We prepared two kinds of libraries: 22 libraries with DNA of a single-species and 6 libraries with a mixture of DNA of several species at known relative concentrations. The single-species libraries were used to calibrate the bioinformatic pipeline of assignment of reads to species; the mixed-species libraries were used to test the ability of the calibrated method to estimate the relative abundance of individual species in mixtures. All libraries were prepared using the TruSeq DNA PCR-Free LT Kit of Illumina following the manufacturer's instructions (Ref. 15037063).

All libraries were sequenced using an Illumina MiSeq with the 2x150 chemistry in three different runs, two runs for single-species (Table 1) and one for mixed-species libraries (Table 2). Four species (*Drosophila melanogaster*, *D. mojavensis*, *D. virilis* and *Linepithema humile*) were sequenced twice in single-species libraries (using different DNA extractions in all cases and different populations for *L. humile* and *D. melanogaster*) to evaluate the repeatability of the method. The mixed-species libraries were prepared from the same extracted DNA of the first run of single-species libraries. Considering the sequencing depth and the genome size of the studied species, the mean coverage obtained was below 1 (Table 1). Therefore, our approach can be qualified as low-coverage shotgun metagenomics.

The target concentration of each species in the mixtures was calculated using a geometric law of parameter k (Magurran 2004): the abundance of the most abundant species is k; the abundance of the second most abundant one is $k \cdot (1-k)$ and so on. The higher the k value, the greater the difference in concentration between species. In the

Table 1. Summary table of the species in single-species libraries; the first run (run #1) was performed in September 2016 and the second (run #2) in July 2018.

Run	Library	Code	Species	Order	Family	Cultured/ Wild	Origin (Country)	Number of raw reads (single-end)	Number of reads after QC and mapping step	Genome coverage
1	1	*PM*	*Papilio machaon*	Lepidoptera	Papilionidae	Wild	Spain	217,260	208,832	0.117
1	2	*DV*	*Drosophila virilis*	Diptera	Drosophilidae	Cultured	Spain	2,357,451	2,088,898	1.717
1	3	*DMe*	*Drosophila melanogaster*	Diptera	Drosophilidae	Cultured	Spain	1,141,884	1,094,403	1.195
1	4	*DMo*	*Drosophila mojavensis*	Diptera	Drosophilidae	Cultured	Spain	834,212	787,688	0.646
1	5	*BO*	*Bactrocera oleae*	Diptera	Tephritidae	Wild	Spain	290,498	279,835	0.093
1	6	*LH*	*Linepithema humile*	Hymenoptera	Formicidae	Wild	Spain	711,171	683,311	0.486
1	7	*AE*	*Acromyrmex echinatior*	Hymenoptera	Formicidae	Cultured	Denmark	116,597	110,086	0.059
1	8	*BT*	*Bombus terrestris*	Hymenoptera	Apidae	Wild	Spain	997,469	972,727	0.603
1	9	*AM*	*Apis mellifera*	Hymenoptera	Apidae	Wild	Spain	631,194	607,965	0.378
1	10	*AP*	*Acyrthosiphon pisum*	Hemiptera	Aphididae	Cultured	USA	342,344	282,940	0.092
2	1	*ACo*	*Atta colombica*	Hymenoptera	Formicidae	Cultured	Denmark	1,636,355	1,607,703	0.845
2	2	*BTa*	*Bemisia tabaci*	Hemiptera	Aleyrodidae	Cultured	Spain	1,256,606	1,170,057	0.307
2	3	*CL*	*Cimex lectularius*	Hemiptera	Cimicidae	Wild	Spain	1,753,361	1,703,285	0.515
2	4	*DMe*	*Drosophila melanogaster*	Diptera	Drosophilidae	Cultured	Spain	1,454,804	1,428,616	1.523
2	5	*DMo*	*Drosophila mojavensis*	Diptera	Drosophilidae	Cultured	Spain	898,735	820,019	0.697
2	6	*DV*	*Drosophila virilis*	Diptera	Drosophilidae	Cultured	Spain	668,442	619,733	0.488
2	7	*DSu*	*Drosophila suzukii*	Diptera	Drosophilidae	Cultured	Spain	1,255,192	1,178,528	0.811
2	8	*LH*	*Linepithema humile*	Hymenoptera	Formicidae	Wild	Spain	1,082,202	1,047,744	0.742
2	9	*PXy*	*Plutella xylostella*	Lepidoptera	Plutellidae	Wild	Spain	2,125,062	1,913,733	0.813
2	10	*SI*	*Solenopsis invicta*	Hymenoptera	Formicidae	Wild	Argentina	1,830,687	1,772,743	0.695
2	11	*VE*	*Vollenhovia emeryi*	Hymenoptera	Formicidae	Wild	Japan	1,743,917	1,679,101	0.912
2	12	*WA*	*Wasmannia auropunctata*	Hymenoptera	Formicidae	Wild	Spain	1,667,606	1,613,371	0.772

Table 2. Relative abundance of the species in mixed-species libraries (in October 2017).

Library	*Acromyrmex echinatior*	*Acyrthosiphon pisum*	*Apis mellifera*	*Bactrocera oleae*	*Bombus terrestris*	*Drosophila melanogaster*	*Drosophila mojavensis*	*Linepithema humile*	*Papilio machaon*	Number of raw reads (single-end)	Number of reads after QC and mapping steps
1	0.5010	0.0078	0.0626	0.2505	0.1252	0.0157	0.0313	0.0039	0.0020	1,897,302	1,842,838
2	0.2505	0.0020	0.1252	0.5010	0.0626	0.0313	0.0157	0.0078	0.0039	1,674,754	1,597,601
3	0.3039	0.0389	0.1088	0.2158	0.1532	0.0548	0.0772	0.0276	0.0196	1,887,006	1,829,291
4	0.2158	0.0196	0.1532	0.3039	0.1088	0.0772	0.0548	0.0389	0.0276	1,557,348	1,511,536
5	0.2127	0.0747	0.1261	0.1787	0.1501	0.0890	0.1059		0.0628	1,767,384	1,709,991
6	0.1787	0.0628	0.1501	0.2127	0.1261	0.1059	0.0890		0.0747	1,344,467	1,264,242

mixtures, we used the following values of k: 0.50 (libraries 1 and 2), 0.30 (3 and 4) and 0.20 (5 and 6). In each library, the order of the species in terms of abundance varied, but several species were only used at low or at high DNA concentrations because of a limitation on the amount of DNA available for each species. Libraries 1–4 contained DNA of 9 species and libraries 5–6 of 8 species (Table 2).

Quality control and mapping

For the sequences generated in the current study, we assessed the quality of raw reads with FastQC v0.11.7 (Andrews 2015). Trimmomatic v0.36 (Bolger et al. 2014) was subsequently used to trim the reads to the specified length of 150 bp and to discard those shorter than 140 bp.

We then aligned each read to all reference genomes individually using BWA 0.7.15-r1140 (Li and Durbin 2009). For each reference, the BWA index was constructed using the index command with default settings. The mapping was conducted with the mem algorithm (Li 2013) and default options. As the mapping of a read was performed independently for each reference, we acquired as many alignment files as references used. We used SAMtools (Li et al. 2009) to remove reads that did not map to any reference. Here, we did not use the paired-end reads provided by the Illumina sequencer, but only the first set of single-end reads (the R1 FASTQ files), because, in many eDNA applications, the fragments were rather short, so the advantage of having paired reads in longer fragments was reduced in actual samples.

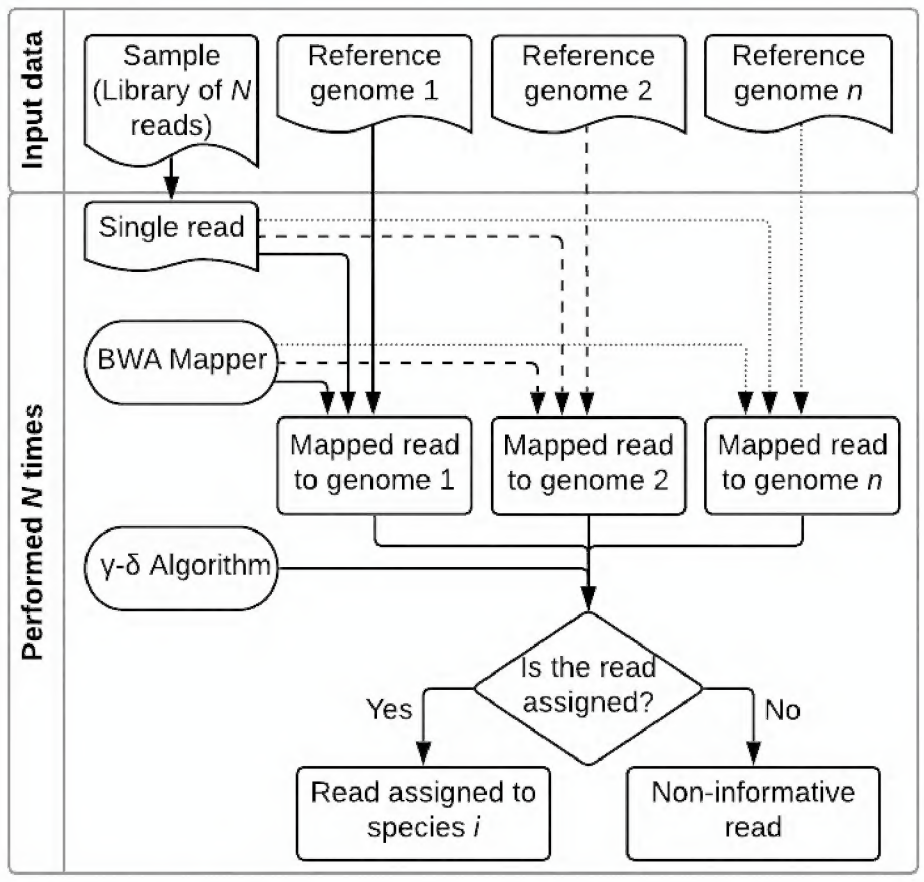

Figure 1. Flow diagram of the computational pipeline used in this study. At the top, input data and, below it, the steps and tools needed for the identification procedure.

From mapped reads to species identity

In general, one-read maps into several reference genomes, so an algorithm is needed to decide between alternative assignations of a read. In metabarcoding and metagenomics studies, reads are commonly assigned to taxa using the lowest common ancestor algorithm (LCA) (MEGAN: Huson et al. 2007; KRAKEN: Wood and Salzberg 2014). The LCA algorithm intends to extract as much taxonomic information as possible from a set of reads, so, if one-read maps well enough in two or more different reference genomes, the LCA assigns the read to their common ancestor in the phylogenetic tree. Here, the interest is different, as we intend to only use genomic regions that are useful for species-level identification; thus, if one-read maps well enough in two different reference genomes, we deem it as non-informative and ignore it, instead of assigning it to its common genus or family. Having this objective in mind, we devised the simple γ–δ algorithm to accomplish it (Figure 2). Basically, what the γ–δ algorithm does is to assign a read to species *i* when it maps well ($\geq \gamma$) to species *i* and bad ($< \delta$) to the rest of species; on the contrary, when a read maps well in two or more species, we declare it non-informative; in all cases $\gamma > \delta$. The quality of the mapping is measured as the mapping ratio *A* and it is calculated as the sum of read's matching nucleotides to the target sequence (n_m), divided by the total number of nucleotides in the alignment (n_t).

$$A = n_m / n_t \quad (1)$$

Even though a read *r* can be assigned to many references, the γ–δ algorithm only needs the two highest mapping ratios. Let A_1 and A_2 be the highest and second highest mapping ratios of *r*. We assume that A_1 corresponds to the mapping ratio of *r* to the reference genome of species *i*. Then the assignation algorithm works in the following way (Figure 2):

- If $A_1 < \gamma$, then *r* is non-informative (because it does not map well enough to any species)
- If $A_1 \geq \gamma$ and $A_2 \geq \delta$, then *r* is non-informative (because it maps too well to two different species)
- If $A_1 \geq \gamma$ and $A_2 < \delta$, then *r* is informative and it is assigned to species *i* (because it maps well enough in one species and not in any other one).

As the best values of γ and δ are unknown, we used the single-species libraries to find the best combination of γ and δ. The reads were divided into a training set (75% of the reads chosen randomly) to find the best γ–δ and a test set (the remaining 25% of reads) to independently calculate the goodness of fit of the model. The tested values of γ and δ were all the combinations of γ = {0.99, 0.98, 0.97} and δ = {0.98, 0.97, 0.96} where γ > δ.

Detection limit

The above algorithm produced a list of species assigned to each read of a library. In single-species libraries, ideally, all reads should belong to the same species (from now on, the focal species). However, detection of additional species could occur for several reasons, such as contamination from the lab, sequencing errors and even tag jumping between multiplexed libraries (Schnell et al. 2015).

In real samples, contaminants are hard to detect, but in our single-species libraries, they are not. If a read corresponds to a species handled simultaneously in the lab, but not sequenced, then it is probably a genuine contamination problem and could be removed from the list of recovered species.

The other kinds of wrongly assigned species likely produce a very low number of reads. The only way to deal with them is to set a detection limit (ε), so the species with a proportion of reads lower than ε are ignored. Here, we present results using the detection limits of 10^{-2}, 10^{-3} and 10^{-4}.

Selection of best values of γ, δ and ε

With the single-species libraries, we used three different criteria to decide which values of γ, δ and ε provided best results. The most important one was that the number of species reported had to be one for single-species libraries. In addition, we wanted to maximise the proportion of reads assigned to the focal species and the proportion of informative reads (assigned to any species).

In practical terms, we first fixed the ε parameter. Next, we compared the different γ–δ combinations using the PERMANOVA test (Anderson 2001), followed by a *post hoc* multiple comparison with the Bonferroni test.

The final output of the above analysis is a combination of values of γ, δ and ε that were best for the single-species libraries analysed in this study. The goodness of fit of this

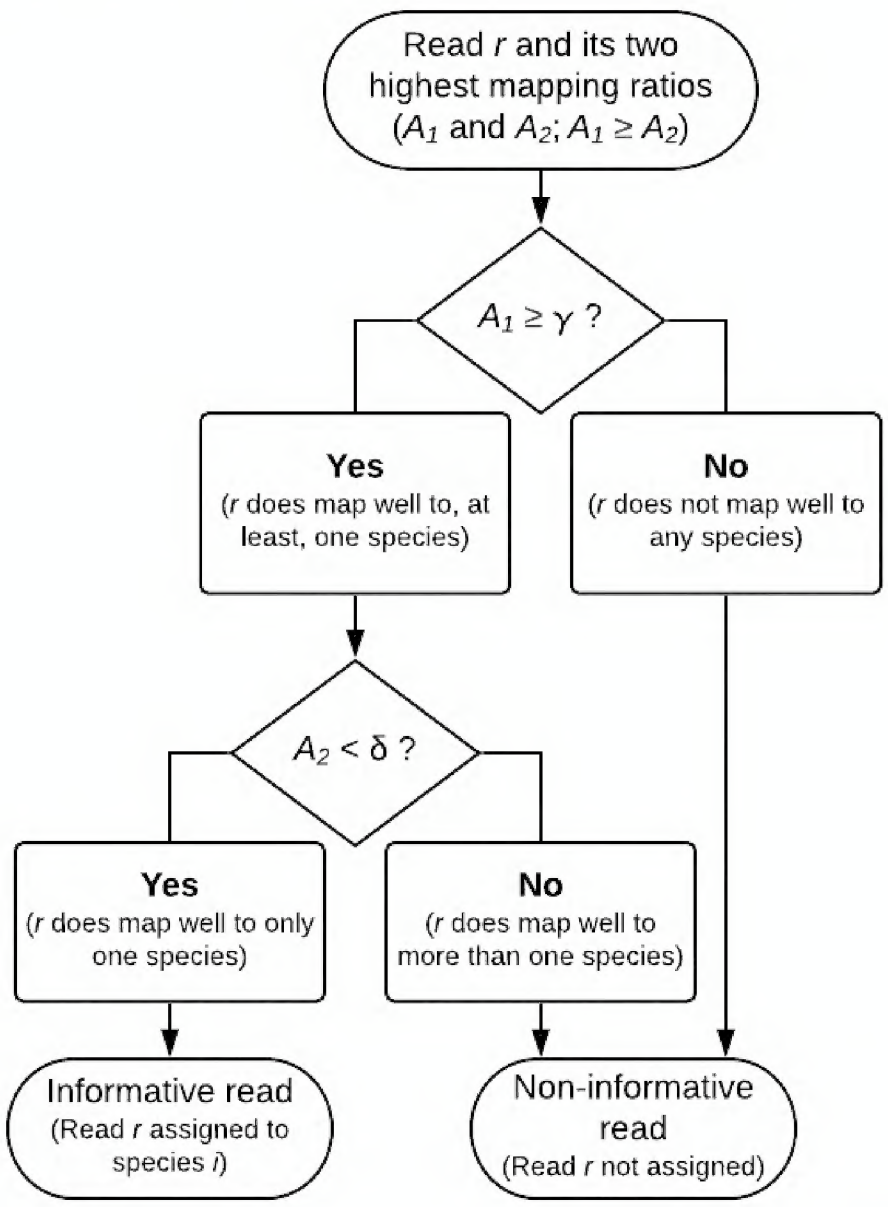
Read r and its two highest mapping ratios (A_1 and A_2; $A_1 \geq A_2$)
$A_1 \geq \gamma$?
Yes
(r does map well to, at least, one species)
No
(r does not map well to any species)
$A_2 < \delta$?
Yes
(r does map well to only one species)
No
(r does map well to more than one species)
Informative read
(Read r assigned to species i)
Non-informative read
(Read r not assigned)

Figure 2. Flow diagram of the γ–δ algorithm. Only the two highest mapping ratios to two reference genomes of a single read are required. In the figure, it is assumed that the highest mapped ratio A_1 belongs to the reference genome *i*.

set of parameters was evaluated using the test set, i.e. the remaining 25% of reads were not used for the calibration.

As will be shown in the results, using the best values of γ, δ and ε, we still found in the single species libraries some reads that were wrongly assigned to non-focal species. To explore the identity of all these misidentified reads, we blasted them (or a subset of 100 reads when the total number of misclassified reads was higher) with megablast (Morgulis et al. 2008) against the NCBI nucleotide collection (nt) database (Wheeler et al. 2007).

Quantification of the relative proportion of the species

Mixed-species mock samples were processed following the same computational pipeline as outlined above (Figures 1 and 2), using the best combination of γ, δ and ε values determined in the previous step. The estimated proportion of reads, assigned to each reference genome, was calculated without considering the rejected reads (not mapped or not assigned reads). This estimated proportion was compared with the actual one (Table 2), using the Pearson correlation coefficient.

Rarefaction of sequenced reads

As can be seen from the results, we obtained a good quantitative estimation of species abundance in all mixed-species libraries. However, from a practical point of view, it would be interesting to investigate if sequencing depth can be reduced and a robust quantitative estimation of the relative species abundance maintained. Thus, more libraries could be multiplexed in a single run and so reduce the overall cost. To evaluate this possibility, we ran the same computational pipeline as before (using the chosen parameters γ, δ and ε), but randomly reducing the number of reads to a proportion of 0.1, 0.01 and 0.001 of the original ones. Each simulation was repeated 100 times, using a different random set of reads. Afterwards, we estimated the number and relative abundance of the recovered species in each rarefied sample and calculated the Pearson correlation between the actual and the estimated relative abundance of the species.

Statistics analysis

All statistical analyses were performed using R 3.4.2 (R Core Team 2016) in RStudio (version 1.0.143; RStudio Team 2015). Permutational analysis of variance (PERMANOVA) and subsequent pairwise comparison with Bonferroni correction were conducted using the package 'vegan' (Oksanen et al. 2018). Plots were created using the packages 'ggplot2' (Wickham 2016) and 'ggpubr' (Kassambara 2018).

Results

Single-species libraries

The 22 libraries prepared from DNA of single-species of insects (Table 1) generated 1,136,957 ± 633,142 (mean ± s.d.) reads, with a coverage of 0.65 ± 0.43. A proportion of 0.013 ± 0.026 reads were eliminated in the trimming step and 0.042 ± 0.026 in the mapping step, so a proportion of 0.95 ± 0.04 of the raw reads remained for further analysis.

The most important characteristic of these libraries is that the number of species recovered, in theory, must be one. With these libraries, we parameterised two aspects of metagenomic species assignment: first, what is the appropriate detection limit (ε) for removal of spurious species (i.e. cut-off for minimum proportion of reads for species to be retained) and second, which are the best values of γ and δ.

For the detection limit ε, we describe in detail the process followed for the analysis of the first run of single-species libraries using the values of γ = 0.99 and δ = 0.98. The rest of the single-species libraries and all the other γ–δ combinations produced relatively similar results and are provided as Supplementary material (Suppl. material 3 and 4).

After the application of the γ–δ algorithm, there were 19.6 ± 8.0 reference genomes (species) per library (Table 3). The most abundant one was the focal species, always above 0.98, except for *B. oleae* (0.93). Obviously, this

Table 3. Proportion of reads assigned to species, in parentheses, for each one of the 10 single-species libraries included in the first sequencing run. The species in each library are shown in the header column. Species assignments are divided in each column into blocks: A, species with abundance higher than ε = 0.01; B, with abundance between ε = 0.01 and ε = 0.001; C, with abundance between ε = 0.001 and ε = 0.0001; D, with abundance below ε = 0.0001; E, potential contaminants. Codes of the species as in Suppl. material 2.

Criteria	Lib. 1 – PM	Lib. 2 – DV	Lib. 3 – DMe	Lib. 4 – DMo	Lib. 5 – BO	Lib. 6 – LH	Lib. 7 – AE	Lib. 8 – BT	Lib. 9 – AM	Lib. 10 – AP
A: above ε = 0.01	**PM** (0.98249)	**DV** (0.9958)	**DMe** (0.99812)	**DMo** (0.99627)	**BO** (0.93349)	**LH** (0.99877)	**AE** (0.99862)	**BT** (0.99777)	**AM** (0.99627)	**AP** (0.99557)
B: from ε = 0.01 to 0.001		LC (0.00259)							AF (0.00226)	
C: from ε = 0.001 to ε = 0.0001			DSi, DSe, DS, DSu, LC	DAr, DEl, DB, LC, DO, DEu, DF, DBi			VE	BI, AF	ACer, AD	MPe
D: below ε = 0.0001	NVi, PP	CQ, NVi, MP, OT, CS, CL, BL, PXy, BI, DAr, DF, DSe, DSu, DT, MDe, DO, NLu	DNo, DY, CL, DW, CQ, DEr, NVi, MP, DF, DO, NLu, DEu	DSi, DT, DSe, DNa, DH, DR, SL, DK, DSu, OT, PXy, DAn	NVi, MDe, LC	BT, CF, SI, CC, TZ	WA, DNo, BI, TS	SI, LC, CCal	EM, DBi	DN, MS, F
E: potential contaminants; **Bold**: species handled in the lab and sequenced; *Italic*: species handled in the lab *but not* sequenced	**DV** (0.01359) *BM* (0.0021) **DMe** (0.00057) **BT** (0.00047) **DMo** (0.0002) **AE** (0.00013) **LH** (0.0001) **AM** (0.00006) **AP** (0.00006) *CCap* (0.00006) **BO** (0.00003)	*BM* (0.00078) *TCa* (0.00025) **DMe** (0.00012) **PM** (0.00011) **LH** (0.00008) **DMo** (0.00003) **BT** (0.00002) **AE** (0.00001) **AM** (0.00001) **BO** (0.00001) **AP** (<0.00001) *CCap* (<0.00001)	*BM* (0.00046) **DV** (0.00031) **LH** (0.00004) **BO** (0.00004) **BT** (0.00003) **DMo** (0.00003) **PM** (0.00002) **AE** (0.00001) **AM** (0.00001) **AP** (0.00001) *TCa* (<0.00001) *CCap* (<0.00001)	*BM* (0.00036) **DV** (0.00008) DMe (0.00004) **LH** (0.00001) **BT** (0.00001) **PM** (0.00001) **AM** (0.00001) **BO** (<0.00001) **AE** (<0.00001) **AP** (<0.00001) *CCap* (<0.00001)	*CCap* (0.0655) *BM* (0.00043) **BT** (0.00015) **DV** (0.00012) **LH** (0.00012) **DMe** (0.00006) **PM** (0.00002) **DMo** (<0.00001)	**AE** (0.0006) *BM* (0.00035) **DV** (0.00005) **DMe** (0.00005) **AP** (0.00003) **AM** (0.00003) **DMo** (0.00002) **BO** (<0.00001) **PM** (<0.00001)	*BM* (0.00055) **LH** (0.00019) **DV** (0.00018) **DMe** (0.00005) **DMo** (0.00005) **BT** (0.00003) **TCo** (0.00003) **BO** (0.00001)	*BM* (0.00078) **DV** (0.00025) **DMe** (0.00024) **DMo** (0.00008) **LH** (0.00004) **BO** (0.00004) **AM** (0.00002) **AP** (0.00002) **PM** (0.00001) **AE** (0.00001) *CCap* (<0.00001)	*BM* (0.00062) **DMe** (0.00014) **DV** (0.00006) **BT** (0.00004) **BO** (0.00004) **LH** (0.00003) **DMo** (0.00002) **PM** (0.00002) **AP** (0.00001) *CCap* (<0.00001)	**DMe** (0.00288) *BM* (0.00071) **LH** (0.00012) **DV** (0.00011) **BT** (0.00009) **DMo** (0.00009) **AM** (0.00003) **BO** (0.00003) **PM** (0.00003) **AE** (0.00001)
Total number of species	14	31	30	32	12	15	14	17	16	15

high number of recovered species is unacceptable for single-species libraries.

Some of these additional species were handled in the same lab, but finally were not sequenced because of their poor quality or for other reasons. These species included *Ceratitis capitata* (Diptera: Trephitidae), *Bombyx mori* (Lepidoptera: Bombycidae) and *Tribolium castaneum* (Coleoptera: Tenebrionidae) in the first single-species run and *B. mori* again in the second run. Thus, they can legitimately appear in the species list because of lab contamination or tag-jumping. If we eliminate these species (Table 3E), the number of species per library is still high at 9.5 ± 11.4 (Tables 3A-D).

The next step is the removal of the species below a certain detection limit. If the species having a relative proportion below $\varepsilon = 0.0001$ were discarded (Table 3D), the remaining number of recovered species would be reduced to 3.1 ± 2.6 (Tables 3A-C). An increase in the detection limit to $\varepsilon = 0.001$ reduced the number of recovered species to one, but for *Drosophila virilis* (*Lucilia cuprina*, same order) and *Apis mellifera* libraries (*Apis florea*, same genus) (Tables 3A-B). A further increase in the detection limit to $\varepsilon = 0.01$ eliminated all non-focal species. In summary, the use of a detection limit of $\varepsilon = 0.001$ almost eliminates all undesired species from the list (Table 3A, B). A very similar result was observed with the single-species libraries of the second run: again, *L. cuprina* appeared in the library of *D. virilis* and *Atta cephalotes* in the library of *A. colombica* (Suppl. material 4). Therefore, considering these results, we will use a detection limit of $\varepsilon = 0.001$ in all further analyses.

The exploration of the misidentified reads in Table 3 against the NCBI nt database produced different kinds of results depending on the species considered. (1) The reads assigned to the dipteran *Lucilia cuprina* in the libraries of the three species of *Drosophila* were assigned to bacteria, mostly *Providencia* sp. and *Morganella* sp. (Suppl. material 7). (2) Ninety-nine percent of the reads assigned to *Apis florea* in the libraries of *Bombus terrestris* and *A. mellifera*, were rRNA and other kinds of RNA. (3) Many reads of *Drosophila melanogaster and D. mojavensis,* assigned to a wrong species of *Drosophila,* mapped into bacteria of the genus *Lactobacillus* and *Acetobacter* and a few were RNAs or transposons. (4) All seven reads of *Acromyrmex echinatior,* wrongly assigned to *Vollenhovia emeryi,* mapped to the bacteria *Wolbachia* sp. (5) Approximately a quarter of the wrongly assigned reads of *Bombus terrestris* to *B. impatiens* were RNAs of *Bombus* or *Apis* and (6) About half of the reads of *Apis mellifera,* assigned to *A. cerana* and *A. dorsata,* were RNAs (Table 3 and Suppl. material 7).

The final step is to decide which of the tested γ–δ combinations provided better results. First, the number of identified species was closer to 1 for $\gamma = 0.99$ than for $\gamma = 0.98$ or $\gamma = 0.97$ (Figure 3A). Even though the combination of $\gamma = 0.98$ and $\delta = 0.96$ was not significantly different from those with $\gamma = 0.99$, that combination had a higher data dispersion of detected species (maximum

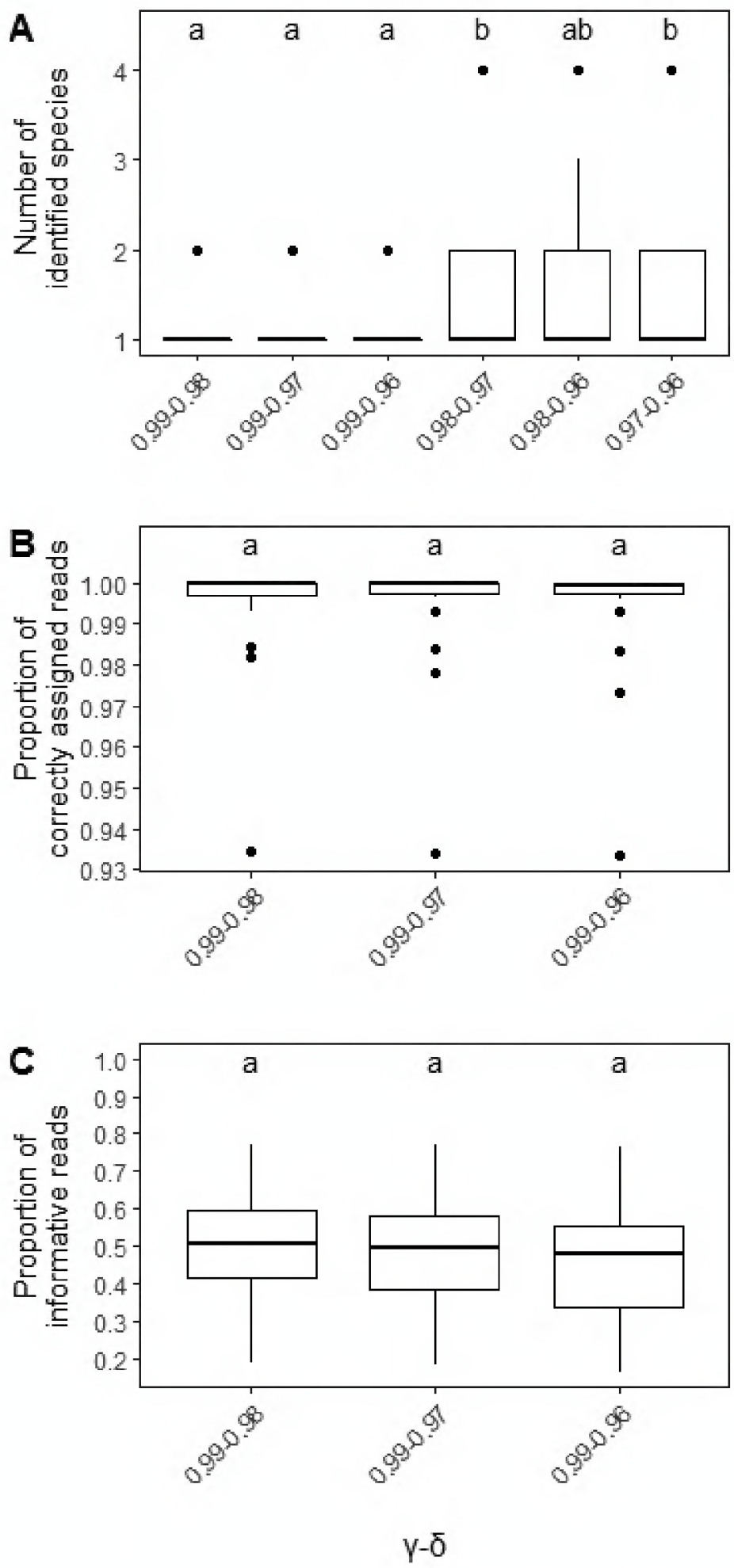

Figure 3. Summary boxplots of the 22 single-species libraries used to search the best combination of parameters γ and δ; in all cases, a detection limit of $\varepsilon = 0.001$ was used and contaminant species were discarded. (**A**) Number of identified species in the library. (**B**) Proportion of the assigned reads allocated to the right species. (**C**) Proportion of the total reads (after trimming and mapping) that were informative (i.e. assigned to any species). A different letter at the top of the figures indicates significant differences amongst γ–δ combinations.

value is 1 vs. 3). Next, we observed neither differences amongst three γ–δ combinations with $\gamma = 0.99$ in the proportion of correctly assigned reads (Figure 3B; $p > 0.99$), nor in the proportion of the informative reads (Figure 3C; $p > 0.91$). Albeit non-significantly, the combination of parameters $\gamma = 0.99$ and $\delta = 0.98$ was slightly better than for

$\delta = 0.97$ or $\delta = 0.96$ and therefore, we will use them in all the following analyses.

The replicated single-species libraries (four species that were analysed in two separate runs) produced remarkably similar results. For example, both libraries of *D. virilis* had *L. cuprina* at a relative concentration higher than $\varepsilon = 0.001$; in the other three libraries, only the focal species was recovered above a value of $\varepsilon = 0.001$ (see Suppl. material 6, for a direct comparison of the duplicated single-species libraries).

We tested the quality of the adjusted parameter set $\varepsilon = 0.001$, $\gamma = 0.99$ and $\delta = 0.98$ obtained with the training set, using the remaining 25% of reads (i.e. the test set). The results were not statistically different between the test set and the training set for any of the analysed variables (Suppl. material 1). Using the test set and the above parameter values, the number of identified species per library was 1.09 ± 0.29, the proportion of correctly assigned reads was 0.99 ± 0.01 and the proportion of informative reads per sample was 0.47 ± 0.15. It is worth noting that the proposed algorithm with the above parameter set was perfectly able to set apart closely-related species, like the species of *Drosophila* (three in the first run and four in the second one).

Mixed-species libraries

The six libraries prepared from DNA of multiple species of insects (Table 2) generated 1,688,044 ± 212,119 reads. A proportion of 0.003 ± 0.001 reads were eliminated in the trimming step and of 0.035 ± 0.012 in the mapping step, so there remained a proportion of 0.962 ± 0.012 of the raw reads for further analysis.

As in the single-species libraries, in the mixed-species libraries, there were also contaminants handled in the laboratory, but not sequenced. To the already mentioned *C. capitata*, *B. mori* and *T. castaneum*, we must add *D. virilis* (that was not sequenced in any mixed-species library) and *L. humile* (not sequenced in libraries 5 and 6). As we did before, we eliminated all these species as genuine contaminants (Table 4E). Even after removing these contaminants, the number of species in the mixture was still very high (45–53), so it was mandatory to apply the proposed detection limit of $\varepsilon = 0.001$ values. By doing this, we recovered all the expected species in the mixtures, 9 in libraries 3–4 and 8 in libraries 5–6; Tables 4A–B), except in libraries 1 and 2 where *P. machaon,* the species with the actual lowest abundance in the mixture, was present in a proportion slightly below $\varepsilon = 0.001$ (Table 4C).

Table 4. Relative proportion of assigned reads to species, in parentheses, for each one of the 6 mixed-species libraries of Table 2 after applying γ–δ algorithm with parameters $\gamma = 0.99$ and $\delta = 0.98$. Codes of the species as in Suppl. material 2. In bold, the species whose DNA was actually put in the mixture.

Criteria	Lib. 1	Lib. 2	Lib. 3	Lib. 4	Lib. 5	Lib. 6
A: above ε = 0.01	**AE** (0.64971) **BO** (0.17167) **BT** (0.06582) **AM** (0.05033) **DMo** (0.02998) **DMe** (0.01052)	**AE** (0.37924) **BO** (0.37805) **AM** (0.12349) **BT** (0.03834) **DMe** (0.02339) **DMo** (0.01707)	**AE** (0.44486) **BO** (0.16181) **AM** (0.10563) **BT** (0.08595) **DMo** (0.0789) **DMe** (0.03823) **LH** (0.0354) **AP** (0.03213)	**AE** (0.30977) **BO** (0.2496) **AM** (0.15797) **BT** (0.06737) **DMo** (0.05985) **DMe** (0.05742) **LH** (0.0536) **AP** (0.01873)	**AE** (0.32342) **BO** (0.1536) **AM** (0.13569) **DMo** (0.12864) **BT** (0.09288) **DMe** (0.0694) **AP** (0.06842) **PM** (0.01576)	**AE** (0.27894) **BO** (0.18874) **AM** (0.16409) **DMo** (0.10473) **DMe** (0.08641) **BT** (0.082) **AP** (0.06055) **PM** (0.01866)
B: from ε = 0.01 to ε = 0.001	**AP** (0.00577) **LH** (0.00245)	**LH** (0.00997) **AP** (0.0017)	**PM** (0.00434)	**PM** (0.00697)		
C: from ε = 0.001 to 0.0001	**PM** (0.00082) VE (0.00013) AF (0.00012)	**PM** (0.0009) AF (0.00026) VE (0.0001)	AF (0.00024) VE (0.00011)	AF (0.00034)	AF (0.00029)	AF (0.00036)
D: below ε = 0.0001	WA, BI, DAr, TCo, ACer, DB, TS, DEl, LC, DO, AD, DBi, TZ, DEu, MDe, CCal, DSi, ACep, CC, DN, DF, DK,DSe, EM, MP, ACo, BD, BL, CL, DR, DSu, NVi, PH, ZC	WA, ACer, BI, LC, AD, DAr, DB, TS, TCo, MDe, DEl, EM, NVi, DO, DSi, DS, DEu, ACo, BL, TZ, ACep, ZC, CQ, DBi, DSe, MP, BD, APl, CF, DT, DW, DY, HL, SI, SC	DAr, WA, DEl, DB, LC, BI, DO, ACer, DBi, MP, TCo, DSi, AD, TS, DEu, DF, MDe, DS, DSe, DT, ACo, TZ, BD, NVi, ACep, SI, DR, BL, DW, CCal, DN, DSu, DNa, DNo, BA, LD, NL, SL	VE, WA, ACer, DAr, DEl, DB, BI, DO, LC, AD, DSi, TS, MP, TCo, DEu, NVi, DF, DSe, MDe, DSu, DY, DBi, DS, ACo, ACep, DNo, CL, PXy, SF, DT, TZ, SI, BL, DW, DN, DNa, CQ, CF, DK, PP	DAr, DEl, VE, DB, DO, WA, LC, BI, ACer, MP, DSi, DF, DBi, DEu, AD, TS, TCo, NVi, DS, TZ, DSe, DW, NL, EM, MDe, DSu, SF, DNa, CQ, PP, ACep, DNo, CL, DT, DN, DK, CCal, BA, LD, APl, CC, DC, DEr	DAr, VE, DEl, ACer, DB, DO, LC, WA, BI, DEu, MP, DSi, DF, DS, DBi, AD, MDe, TCo, DSu, TS, DSe, EM, DNo, NVi, NL, DNa, ACep, DT, ACo, DH, TZ, CQ, PP, CL, DN, ZC, HL, AGa, Dan, MS, PR
E: potential contaminants	CCap (0.01185) BM (0.0005)	CCap (0.02603) BM (0.00112) TCa (< 0.00001)	CCap (0.01131) BM (0.00051) TCa (< 0.00001) DV (< 0.00001)	CCap (0.01743) BM (0.00044)	CCap (0.01064) BM (0.00055) LH (< 0.00001)	CCap (0.01323) BM (0.00155) LH (< 0.00001) DV (< 0.00001)
Total number of species	47	49	53	52	55	54

The correlation coefficient between actual and estimated relative species abundances was statistically significant in all mixtures (Figure 4), so the method was able to quantify the relative proportions of the species. The fitting was better for high values of k (more difference in the relative abundance of species; libraries 1–2, $k = 0.50$) than for low values of k (less difference in the relative abundance; libraries 5–6, $k = 0.20$) (Figure 4).

We run the entire pipeline on a server with 2 Intel Xeon E5-2620 v3 processors with 6 cores each, which allowed a maximum of 24 threads, thanks to their hyper-threading technology. The total processing time varied between 54 min (library #6, 1.3 raw million reads) and 1 h 16 min (library #1, 1.9 raw million reads), most of it (89%) consumed by the mapping of the reads into the reference genomes and very little (3–4%) by the γ–δ algorithm (see Suppl. material 8 for the processing time of each step of the pipeline for 6 mixed-species libraries).

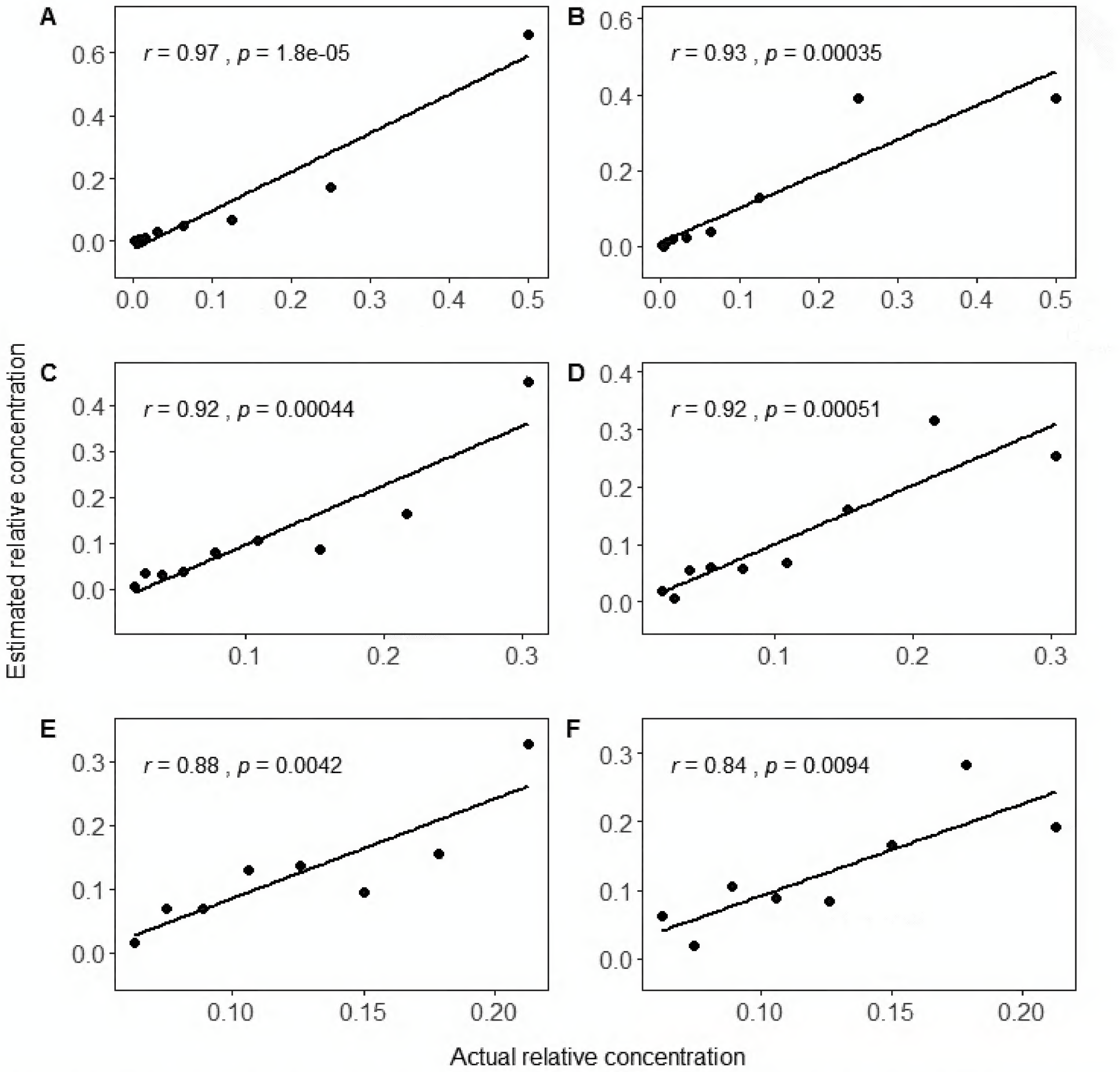

Figure 4. Scatter plots between the expected (i.e. as the mixtures were prepared in the lab; Table 2) and the estimated species relative abundance following the described bioinformatic pipeline (Table 4). Each plot corresponds to one mixed-species library (A to F corresponds to libraries 1 to 6). Each point in the plot indicates one species in the mixture. In each plot, the correlation coefficient (r) and its p-value are also indicated.

Rarefaction of the reads

When only a proportion of 0.1 or even 0.01 of the initial reads was used, the number of recovered species was the same in libraries 3–6 as when all reads were used (Figures 5C-F). In libraries 1 and 2, there was some discrepancy, but it was caused by the estimated relative abundance of *P. machaon* being sometimes slightly below and sometimes slightly above 0.001 and so our detection limit of $\varepsilon = 0.001$ discarded or accepted the species accordingly (Figures 5A-B). A further reduction in the proportion of used reads to 0.001 made the number of identified species less predictable (Figure 5). However, the correlation coefficient *r* between the observed and the expected relative abundance was always significant at all rarefaction levels (Figure 6).

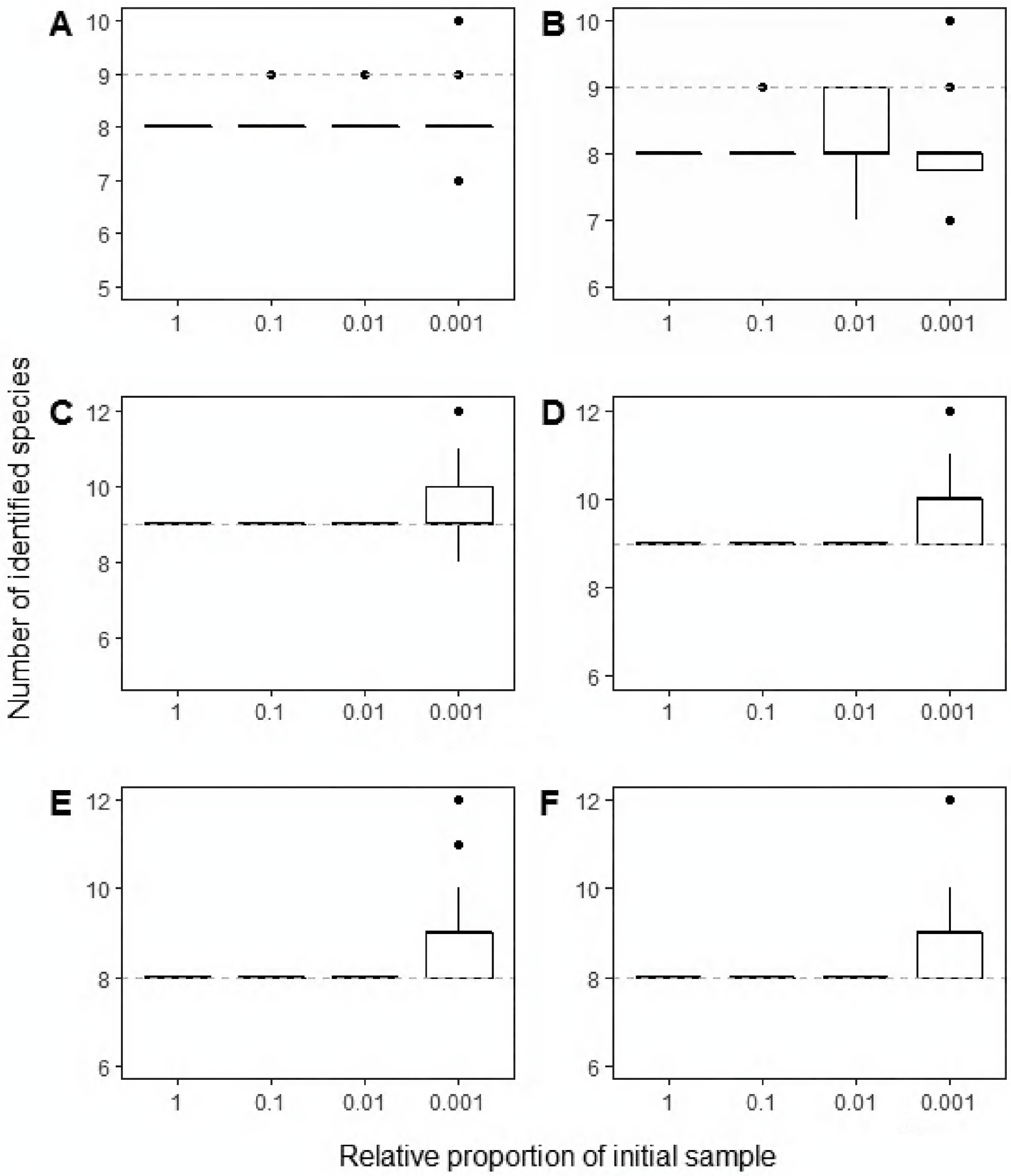

Figure 5. Effect of the rarefaction of reads on the number of species detected (above $\varepsilon = 0.001$ and without contaminants) in the six mixed-species libraries (A to F correspond to libraries 1 to 6). The x axis indicates the proportion of reads used (when 1, all reads were used, so there is only one value); in the rest of the values, 100 random repetitions were conducted using the indicated proportion of reads. The horizontal dashed line of each plot indicates the actual number of species in the mixture.

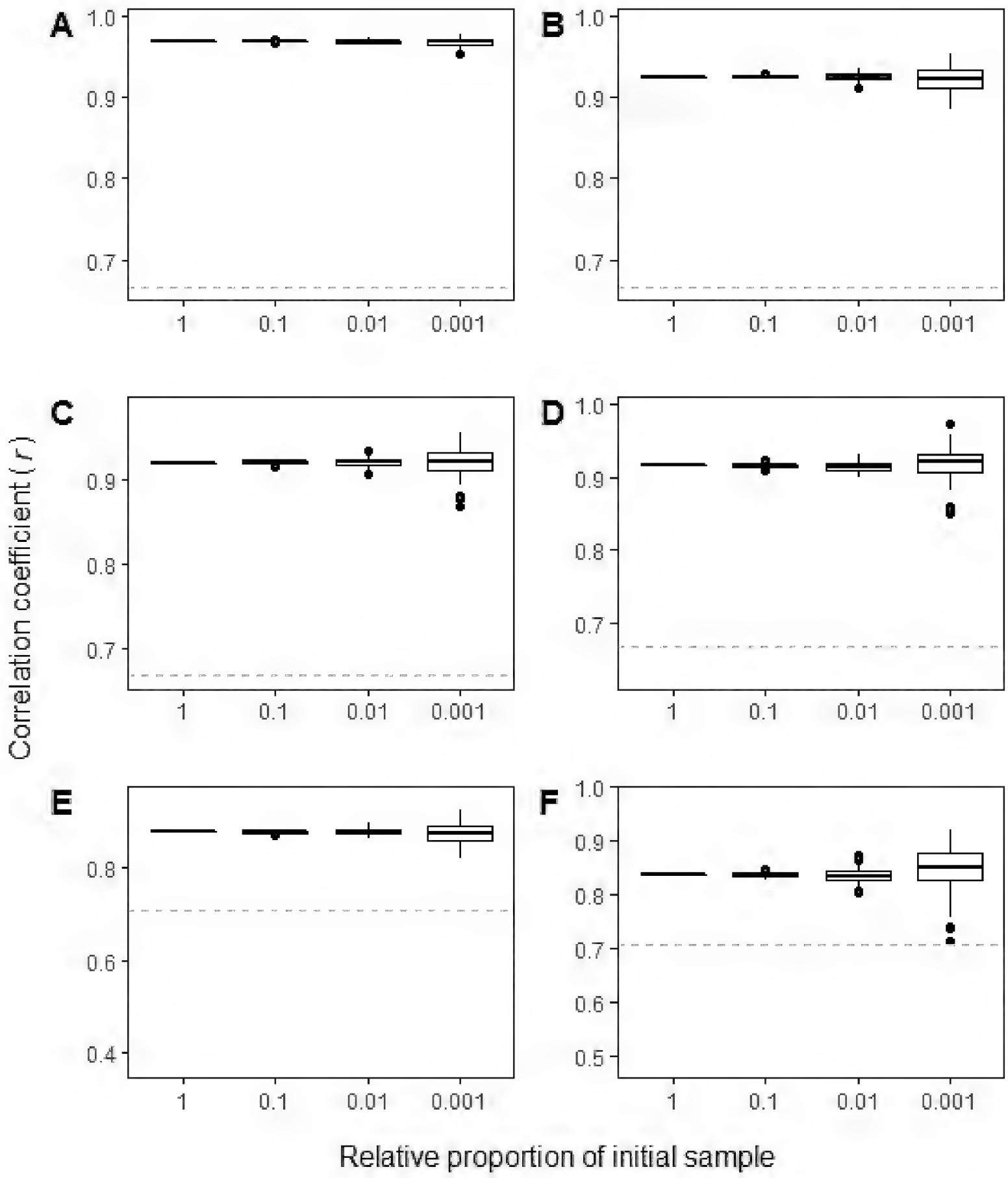

Figure 6. Effect of the rarefaction of reads on the correlation coefficient *r* between the expected and the recovered relative abundance of the species in the six mixed-species libraries (A to F correspond to libraries 1 to 6). The x axis indicates the proportion of reads used (when 1, all reads were used, so there is only one value); in the rest of the values, 100 random repetitions were conducted using the indicated proportion of reads. The horizontal dashed line of each plot indicates the critical value of *r*, above which measured *r* is statistically significant at $p < 0.05$.

Discussion

Metagenomics is a technology devised to obtain both taxonomic and functional gene information for entire communities of organisms (Thomas et al. 2012; Zepeda Mendoza et al. 2015) and its use is more common in prokaryotes than in eukaryotes. Here, we focused on the taxonomic aspect of metagenomics and applied it to Metazoa. We evaluated the technique using artificial mixtures of DNA consisting from one to nine insect species whose complete genome has been sequenced to an advanced degree. The single-species libraries proved to be very useful in showing the limitations of the technique: in these libraries, the number of expected species is one, but we found between 12 and 32 species per library, so it was mandatory to establish a detection limit for a species to be included in the species list. The mixed-species libraries showed that the technique is perfectly able to quantitatively determine the relative abundance of individual species in mixtures. Given the scarcity of assembled genomes of Metazoa, the proposed methodology is a proof of concept of the metagenomics approach rather than a method to be applied immediately to actual environmental samples.

Species identification: spurious species and the need for an analytical limit of detection

Our data, collected from single insect specimens, produced assignments to ca. 20 species; similarly, in each mixed-species library (8–9 species), ca. 50 species were recovered, so most of the listed species were spurious (Tables 3 and 4). These extra species can be divided into two groups; contaminants (species handled simultaneously in the same lab) and species for which there is no known reason for their presence.

There are two possible causes for contaminant DNA. The first one is physical contamination in the preparation of the libraries in the lab; the second one is the index-hopping effect during the sequencing reaction (Schnell et al. 2015). We had examples of both kinds.

Three species were handled simultaneously in the lab, but not sequenced. All these species appear in most libraries, generally in a proportion lower than 0.001 (Tables 3 and 4 and Suppl. material 4). Most of the reported contaminants cannot be attributed to specific issues in the lab workflow. However, the presence of *Ceratitis capitata* in libraries of *Bractocera oleae* (it accounts for 6.6% in the single-species library and above 1% in the mixed-species ones) may have occurred during sample collection. These two dipterans were trapped together in agricultural fields and also transported together to the lab. There, a trained entomologist separated the individuals of the two species; it is very unlikely that this person could have made an identification mistake, but it is possible that fragments of *C. capitata* (legs, antennae, wings) ended up in the *B. oleae* tube. In addition, the two species were, for a certain period, suspended in the same ethanol solution. In the analysis of our artificial mock samples (both single and mixed-species), we eliminated all the contaminant species because we knew that they were contaminants. However, in actual environmental samples, it can be challenging to set apart contaminants from species belonging to the community.

Another possibility for inter-sample contamination is the worrisome tag-jumping effect (Schnell et al. 2015), in which a read from one library is mistakenly taken as belonging to another one because the tag, used to identify each multiplexed library, is sequenced erroneously. Of course, it is not possible to distinguish this process from the genuine contamination discussed above. Again, we could safely ignore these species in the single-species libraries, but we cannot do anything about them in the mixed-species libraries nor in the real samples.

In addition to the contaminants, many other species appeared in the lists of both the single and mixed-species libraries (Tables 3 and 4) that were never handled in our lab nor could be found in the area. All these species appeared at small relative abundances, almost always below the threshold of $\varepsilon = 0.001$. The cause of these misclassifications is probably a sequencing error in our samples, but there must also be errors and missing sequences in the reference genomes themselves (Donovan et al. 2018; Lu and Salzberg 2018). For example, some of the wrongly assigned reads were of mutualistic or parasitic bacteria of insects, like *Providencia* sp., *Morganella* sp., *Lactobacillus* sp., *Acetobacter* sp. and *Wolbachia* sp. (Chandler et al. 2011; Singh et al. 2015; Simhadri et al. 2017). Thus, it is reasonable to assume that they were in our samples alongside the insects, but also in the specimens used to generate the reference genomes. Several other wrongly assigned reads were of conserved RNA sequences that are difficult to set apart from phylogenetically similar species. In addition, there is always some intraspecific genetic variability in all species and the specimens that we sequenced likely come from a different population from the one used to obtain the reference genome.

The only way to eliminate these species from the species list of each library is to set a threshold for the relative abundance of the species, i.e. an analytical detection limit. A detection limit of $\varepsilon = 0.001$ eliminated all the unwanted species in all but 3 of the 28 artificial libraries (Tables 3 and 4 and Suppl. material 4). There is a reasonable explanation for two of these three misplaced species, as they were congeneric species in the honey bee *Apis mellifera* (*A. florea*) and in *Atta colombica* (*A. cephalonica*) libraries. The presence of the Dipteran *Lucilia cuprina* in the two libraries of *Drosophila virilis* seems to be mediated by two bacteria (*Providencia* sp. and *Morganella* sp.) associated with the microbiome of dipterans (Chandler et al. 2011; Singh et al. 2015) that appear in the published genome of *L. cuprina*. Merchant et al. (2014) show that this problem is widespread, as they found bacterial contamination in five out of nine eukaryotic assembled and published genomes. New bioinformatic tools for the decontamination of eukaryotic genome assemblies from bacterial contaminants (Fierst and Murdock 2017) are likely to alleviate this problem.

In the mixed-species libraries, the detection limit of $\varepsilon = 0.001$ removed all spurious species, with no exceptions. However, in libraries 1 and 2, DNA of *Papilio machaon* was used to prepare the mixtures at a low concentration (Table 2) but was excluded from the species list (Table 4). Therefore, on one hand, the use of a detection limit has the desired effect of eliminating false positives but, on the other hand, can generate false negatives. In our mixed-species libraries, the balance was favourable, as there were no false positives and only two false negatives.

We do not think that the presence of spurious species in our artificial libraries is specific to the way that we handled the DNA in the lab or to our species assignation algorithm. The problem is probably more general, but it is only exposed when artificial samples are analysed, especially in those consisting of only one species. Other researchers have found similar results using prokaryotes (Pereira et al. 2018). Consequently, we recommend the use of a stringent detection limit (e.g. $\varepsilon = 0.001$) to avoid a long list of spurious species. Of course, this will have the negative effect of excluding some species that actually are present at low abundance, but this trade-off between false positives and false negatives is inevitable (Alberdi et al. 2017). To be fair, most studies already do this but in

a rather unsystematic way. For instance, MEGAN (Huson et al. 2007) and many other studies always ignore singletons. Other studies increase the minimum number of reads to keep a taxon in the list (five in Piñol et al. 2014; ten in Gibson et al. 2015 and in Lee et al. 2018). As we do here, Pompanon et al. (2012) and Alberdi et al. (2017) suggest that a relative threshold can be more appropriate than absolute read count thresholds.

Quantification of the relative abundance of the species

The main objective of using metagenomics for the quantification of the species abundance and hence of this study, was to overcome the PCR-biases of amplicon metabarcoding. Here, we showed that the metagenomics approach completely fulfilled this objective, whereas in amplicon metabarcoding, the quantification of the abundances of the species is sometimes good (Saitoh et al. 2016; Kraaijeveld et al. 2015), but in others, it is very poor (Piñol et al. 2015; Leray and Knowlton 2017).

Our mixed-species libraries comprised ca. 1.7 million reads each, but the rarefaction experiment showed that, even with 100 times less reads (ca. 17000 per library), the quantification would still be good (Figure 6). Thus, many more samples could be multiplexed in one single Illumina MiSeq run and, consequently, reduce the mean cost per library. Of course, if the mixtures were richer in species, more reads per sample would be needed. Greenwald et al. (2017) applied shotgun metagenomics in prokaryotes and was also able to estimate relative species abundance with high fidelity ($r^2 > 0.92$).

However, it is important to remember that not all biases are corrected by shotgun metagenomics. Here, we began the process using extracted DNA, so all the biases in the generation of eDNA sequences (i.e. digestion rates in dietary studies or DNA degradation in the soil or in the water, or in the DNA extraction) are not accounted for. In particular, the same amount of biomass does not always render the same amount of DNA (Pornon et al. 2016); thus, as the usual goal is the estimation of species biomass, a biomass-to-DNA factor should be estimated for each species or, alternatively, the artificial mixtures should be prepared from a known biomass of each species rather than from a known DNA amount, as some authors already do (Zhou et al. 2013; Tang et al. 2015).

Data treatment and the assignation of reads to species

In metagenomics, there are, basically, two methods to assign reads to species; the assembly-based and the read-based approaches (Thomas et al. 2012). In the former, the reads are assembled using a de-novo assembler into contigs and these are mapped into reference genomes; the quantification of the species is achieved by counting the number of reads assembled in contigs that map into a given species. This approach is commonly used in prokaryote and in mitochondrial metagenomics, but it was not useful in this application because of the low coverage of our sequencing: with so few overlapping reads, many very small contigs would be obtained.

Consequently, we used here the read-based approach that assigns a species to every read by mapping it into a reference genome. As a mapper, we used BWA, but other possibilities would probably be good choices too (e.g. Bowtie2, MagicBlast, GEM; Langmead and Salzberg 2012; Boratyn et al. 2018; Marco-Sola et al. 2012). In any case, all mappers normally produce hits of one read into several reference genomes, so an algorithm is needed to assign a species to a read. By far the most common algorithm used in metabarcoding and metagenomics studies is the lowest common ancestor algorithm (LCA; MEGAN: Huson et al. 2007; KRAKEN: Wood and Salzberg 2014); albeit, there are other alternatives (Hanson et al. 2016; Sarmashghi et al. 2019). However, we used here our own γ–δ algorithm that sets species apart rather than extracting as much taxonomic information as possible from a set of reads, as the LCA algorithm does. The γ–δ algorithm declared, as informative, approximately half of the reads. This algorithm is extremely straightforward and easy to implement.

Present and future of metagenomics

The metagenomics approach presented here for eukaryotic species will not be a realistic option until the number of sequenced genomes is a substantial fraction of the total biodiversity. Today, the metagenomic method for taxonomic purposes is used mostly with genomes of organelles instead of whole genomes, because the number of sequenced organelle genomes is much higher than the number of whole genomes (e.g. today there are roughly, in the NCBI RefSeq database, 14 times more mitogenomes than whole genomes of insects). In addition, the number of sequenced organelle genomes is increasing quickly with new easier and faster methods, based on next generation sequencing and de novo assembly (Cameron 2014).

Mito-metagenomics has proved to be better than amplicon metabarcoding for quantification purposes (Gómez-Rodríguez et al. 2015; Bista et al. 2018; Gueuning et al. 2019), but estimation of relative abundance amongst species (i.e. in a given sample, species A is more abundant than species B) is not always high (Tang et al. 2015; Krehenwinkel et al. 2017). The quantification power of mito-metagenomics is likely bounded for two reasons. First, when there is no mitochondrion enrichment (as in Zhou et al. 2013), only a small proportion of the shotgun reads map into the mitogenome (~0.5 % in insects; Tang et al. 2014), so a high sequencing depth is necessary to obtain good quantitative results (Gueuning et al. 2019). Second and most important, the number of mitogenomes per nuclear genome (mitochondrial copy number) is variable amongst species and even between tissues. Consequently, in a given amount of DNA (and using it as a proxy of biomass), the mitochondrial copy number will vary across species, so the estimation of the relative abundance of the species will be affected. This problem

is known (Crampton-Platt et al. 2016; Krehenwinkel et al. 2017) and applies not only to mito-metagenomics, but also to amplicon metabarcoding targeting mitochondrial markers. The solution is to use an independent estimation of the mitochondrial copy number for each species, but we are not aware of any reliable data of this variable across arthropod species.

It is also fair to question about the computational problems that would pose a future with huge reference databases when the genomes of most species are sequenced. Perhaps our implementation of the method could be so computationally costly that it would be inapplicable in practice. In our opinion, the method is perfectly manageable today in a modest computer server and will remain so in the foreseeable future. In the reported experiments, the maximum processing time of the entire pipeline per mixed-species library was of 1.3 hours, most of it being devoted to the mapping of reads into the reference genomes. This mapping of reads into genomes is a problem that fits into the category known as "embarrassingly parallel" applications (McCool et al. 2012), in which a read can be processed simultaneously with different references and, therefore, the complexity of the algorithm increases linearly with the number of reads n and the number of genomes g. Thus, using library #1 as an example, multiplying g by 100 (~ 11000 genomes in our case) and decreasing n by 100 (~ 19000 reads) should keep the execution time roughly at the same 1.3 hours (we showed here that it was possible to reduce the number of reads without loss of identification and quantification power; Figures 5 and 6).

In addition, the pipeline could eventually be modified in several ways to further reduce the execution time. (1) Selection of reference genomes in the database: when processing a sample, there is no need to compare the reads with all the animal genomes (or plant or fungi) in the world: if the interest is in insects, then only the genomes of insects known to occur in a certain geographical region should become the reference database. Thus, even in a future with the genomes of *all* species already sequenced, the number of genomes of interest will never be of millions, but of 10^3 to 10^5 genomes at most. (2) Filtering of the reference genomes database: we showed here that only approximately half of the reads were informative. The non-informative reads probably belong to certain regions of the genome that, when identified, could be filtered out from the reference database with appropriate programmes. (3) Elimination of non-informative reads in running time: in the γ–δ algorithm, a read that maps better than δ in two different genomes is declared non-informative. Once that read is detected, the mapping of it against the remaining reference genomes is not necessary anymore and finally (4) It is reasonable to assume that the power of the computers will continue to increase in the future as it has done in the past (Williams 2017). It is even possible that, in the next decades, unimagined computational capabilities become available with the advent of quantic processors (Arute et al. 2019).

Concluding remarks

According to our results, the low-coverage shotgun metagenomic method is perfectly capable to set apart closely related insect species, like the four species of the genus *Drosophila* that we included in the artificial libraries. We also saw that, despite the risk that some reads were not in the reference databases that we used (reads of commensal or parasites species; parts of the genome not yet sequenced) or that some reads were very similar in more than one reference genome, we achieved a reasonable proportion (ca. 0.50) of truly informative reads. By using mixtures, we showed that it is possible with this technique to quantify with confidence the relative abundance of individual species in the mixtures and that, with much less sequencing depth than the one used here, it was possible to obtain comparable results (ca. 17000 reads in mixtures of ca. 10 species). Finally, a word of caution. The "dream" of getting an eDNA sample, sequencing it, mapping it against a growing DNA database and obtaining the species names and relative abundance of all species in the mixture that we tried to simulate in this study, is not without hurdles. The main one is obviously the low number and quality of eukaryote genomes sequenced so far, but also the impossibility of identifying, with confidence, species below a certain detection limit and the need to improve the algorithms in a future with huge genome databases and increased sequencing depth.

Data accessibility

The data underpinning the analysis reported in this paper are deposited in the Dryad data Repository at https://doi.org/10.5061/dryad.t1g1jwsz7. The entire pipeline and the γ–δ algorithm implementation are available at GitHub: https://github.com/LidiaGS/g-d_algorithm.

Authors' contributions

The experiment was conceived by JP and designed by all authors. Bioinformatic analyses were performed by LGS and supervised by the other two authors. All authors contributed to the writing of the manuscript.

Acknowledgements

We are very grateful to several entomologists who kindly provided specimens for sequencing: Xavier Espadaler, Nicolás Pérez, Alfredo Ruíz, Francesc Mestres, Aleix Valls, Francisco Beitia, Carlos Hernández-Castellano, Joan Josep Ibañez, Carlos Pradera, Rasmus S. Larsen, Jacobus J. Boomsma, Luis Calcaterra and Misato O. Miyakawa. We also thank Anna Barceló and Roger Lahoz of the Genomics facilities of the UAB for extracting and sequencing the DNA. The comments of Bruce E. Deagle, Alfried P. Vogler, Yiyuan Li and Xin Zhou on earlier

versions of the manuscript are also much appreciated. Financial help was provided by the grants TIN2017-84553-C2-1-R (Spanish Government) and AGAUR 2017 SGR 1001 (Generalitat de Catalunya).

References

Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K (2017) Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution 9: 134–147. https://doi.org/10.1111/2041-210X.12849

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. Austral Ecology 26: 32–46. http://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Andrews S (2015) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arute F, Arya K, Babbush R, Bacon D, Bardin JC, Barends R, Biswas R, Boixo S, Brandao FGSL, Buell DA, Burkett B, Chen Y, Chen Z, Chiaro B, Collins R, Courtney W, Dunsworth A, Farhi E, Foxen B, Fowler A, Gidney C, Giustina M, Graff R, Guerin K, Habegger S, Harrigan MP, Hartmann MJ, Ho A, Hoffmann M, Huang T, Humble TS, Isakov SV, Jeffrey E, Jiang Z, Kafri D, Kechedzhi K, Kelly J, Klimov PV, Knysh S, Korotkov A, Kostritsa F, Landhuis D, Lindmark M, Lucero E, Lyakh D, Mandrà S, McClean JR, McEwen M, Megrant A, Mi X, Michielsen K, Mohseni M, Mutus J, Naaman O, Neeley M, Neill C, Niu MY, Ostby E, Petukhov A, Platt JC, Quintana C, Rieffel EG, Roushan P, Rubin NC, Sank D, Satzinger KJ, Smelyanskiy V, Sung KJ, Trevithick MD, Vainsencher A, Villalonga B, White T, Yao ZJ, Yeh P, Zalcman A, Neven H, Martinis JM (2019) Quantum supremacy using a programmable superconducting processor. Nature 574: 505–510. https://doi.org/10.1038/s41586-019-1666-5

Bista I, Carvalho GR, Tang M, Walsh K, Zhou X, Hajibabaei M, Shokralla S, Seymour M, Bradley D, Liu S, Christmas M, Creer S (2018) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. Molecular Ecology Resources 18: 1020–1034. https://doi.org/10.1111/1755-0998.12888

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15): 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boomsma JJ, Brady SáG, Dunn RR, Gadau J, Heinze J, Keller L, Sanders NJ, Schrader L, Schultz TR, Sundström L, Ward PS, Wcislo WT, Zhang G, The GAGA Consortium (2017) The Global Ant Genomics Alliance (GAGA). Myrmecological News 25: 61–66.

Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL (2018) Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. bioRxiv. https://doi.org/10.1101/390013

Cameron SL (2014) How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research. Systematic Entomology 39(3): 400–411. https://doi.org/10.1111/syen.12071

Chandler JA, Morgan Lang J, Bhatnagar S, Eisen JA, Kopp A (2011) Bacterial Communities of Diverse *Drosophila* Species: Ecological Context of a Host-Microbe Model System. PLoS Genetics 7(9): e1002272. https://doi.org/10.1371/journal.pgen.1002272

Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux P-M, Li F-W, Melkonian B, Mavrodiev EV, Sun W, Fu Y, Yang H, Soltis DE, Graham SW, Soltis PS, Liu X, Xu X, Wong GK-S (2018) 10KP: A phylodiverse genome sequencing plan. GigaScience 7(3): giy013. https://doi.org/10.1093/gigascience/giy013

Chen K, Pachter L (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Computational Biology 1(2): e24. http://doi.org/10.1371/journal.pcbi.0010024

Clare EL, Chain FJJ, Littlefair JE, Cristescu ME (2016) The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. Genome 59(11): 981–990. http://doi.org/10.1139/gen-2015-0184

Crampton-Platt A, Yu DW, Zhou X, Vogler AP (2016) Mitochondrial metagenomics: letting the genes out of the bottle. Gigascience 5(1): 15. https://doi.org/10.1186/s13742-016-0120-y

Deagle BE, Thomas AC, McInnes JC, Clarke LJ, Vesterinen EJ, Clare EL, Kartzinel TR, Eveson JP (2018) Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data?. Molecular Ecology 28: 391–406. http://doi.org/10.1111/mec.14734

Donovan PD, Gonzalez G, Higgins DG, Butler G, Ito K (2018) Identification of fungi in shotgun metagenomics datasets. PLoS ONE 13(2): e0192898. https://doi.org/10.1371/journal.pone.0192898

Elbrecht V, Leese F (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. PLoS ONE 10(7): e0130324. https://doi.org/10.1371/journal.pone.0130324

Evans NT, Olds BP, Renshaw MA, Turner CR, Li Y, Jerde CL, Mahon AR, Pfrender ME, Lamberti GA, Lodge DM (2016) Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. Molecular Ecology Resources 16(1): 29–41. https://doi.org/10.1111/1755-0998.12433

Fierst J, Murdock DA (2017) Decontaminating eukaryotic genome assemblies with machine learning. BMC Bioinformatics 18:533. https://doi.org/10.1186/s12859-017-1941-0

Hanson NW, Konwar KM, Hallam SJ (2016) LCA*: an entropy-based measure for taxonomic assignment within assembled metagenomes. Bioinformatics 32(23): 3535–3542. https://doi.org/10.1093/bioinformatics/btw400

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Research 17: 377–386. https://doi.org/10.1101/gr.5969107

i5K Consortium (2013) The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. Journal of Heredity 104 (5): 595–600. https://doi.org/10.1093/jhered/est050

Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. Journal of Heredity 100(6): 659–674. http://doi.org/10.1093/jhered/esp086

Gibson JF, Shokralla S, Curry C, Baird DJ, Monk WA, King I, Hajibabaei M (2015) Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. PLoS ONE 10(10): e0138432. https://doi.org/10.1371/journal.pone.0138432

GIGA Community of Scientists (2014) The global invertebrate genomics alliance (GIGA): Developing community resources to study diverse invertebrate genomes. Journal of Heredity 105(1): 1–18. https://doi.org/10.1093/jhered/est084

Gómez-Rodríguez C, Crampton-Platt A, Timmermans MJTN, Baselga A, Vogler AP (2015) Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. Methods in Ecology and Evolution 6(8): 883–894. https://doi.org/10.1111/2041-210X.12376

Greenwald WW, Klitgord N, Seguritan V, Yooseph S, Venter JC, Garner C, Nelson KE, Li W (2017) Utilization of defined microbial communities enables effective evaluation of meta-genomic assemblies. BMC Genomics 18: 296. https://doi.org/10.1186/s12864-017-3679-5

Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I (2014) MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Research 42(D1): D699–D704. https://doi.org/10.1093/nar/gkt1183

Gueuning M, Ganser D, Blaser S, Albrecht M, Knop E, Praz C, Frey JE (2019) Evaluating next-generation sequencing (NGS) methods for routine monitoring of wild bees: Metabarcoding, mitogenomics or NGS barcoding. Molecular Ecology Resources 19(4): 847–862. https://doi.org/10.1111/1755-0998.13013

Kassambara A (2018) ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.2. https://CRAN.R-project.org/package=ggpubr

Kraaijeveld K, de Weger LA, Ventayol García M, Buermans H, Frank J, Hiemstra PS, den Dunnen JT (2015) Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. Molecular Ecology Resources 15(1): 8–16. https://doi.org/10.1111/1755-0998.12288

Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG (2017) Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports 7: 17668. http://doi.org/10.1038/s41598-017-17333-x

Lamb PD, Hunter E, Pinnegar JK, Creer S, Davies RG, Taylor MI (2018) How quantitative is metabarcoding: A meta-analytical approach. Molecular Ecology 28(2): 420–430. http://doi.org/10.1111/mec.14920

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359. https://doi.org/10.1038/nmeth.1923

Lee TR, Alemseged Y, Mitchell A (2018) Dropping Hints: Estimating the diets of livestock in rangelands using DNA metabarcoding of faeces. Metabarcoding and Metagenomics 2: 1–17. http://doi.org/10.3897/mbmg.2.22467

Leray M, Knowlton N (2017) Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. PeerJ 5: e3006. http://doi.org/10.7717/peerj.3006

Levine R (2011) i5k: the 5,000 insect genome project. American Entomologist 57(2): 110–113. https://doi.org/10.1093/ae/57.2.110

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, Goldstein MM, Grigoriev IV, Hackett KJ, Haussler D, Jarvis ED, Johnson WE, Patrinos A, Richards S, Castilla-Rubio JC, van Sluys M-A, Soltis PS, Xu X, Yang H, Zhang G (2018) Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences 115(17): 4325–4333. https://doi.org/10.1073/pnas.1720115115

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14): 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16): 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Linard B, Crampton-Platt A, Cillett CPDT, Timmermans MJTN, Vogler AP (2015) Metagenome skimming of insect specimen pools: potential for comparative genomics. Genome Biology and Evolution 7(6): 1474–1489. https://doi.org/10.1093/gbe/evv086

Lu J, Salzberg SL (2018) Removing contaminants from metagenomic databases. PLoS Computational Biology 14(6): e1006277. https://doi.org/10.1371/journal.pcbi.1006277

Magurran AE (2004) Measuring Biological Diversity. The Journal of the Torrey Botanical Society 131(3): 277–278. http://doi.org/10.2307/4126959

Marco-Sola S, Sammeth M, Guigó R, Ribeca P (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. Nature Methods 9: 1185–1188. https://doi.org/10.1038/nmeth.2221

McCool M, Reinders J, Robison A (2012) Structured Parallel Programming: Patterns for Efficient Computation. Morgan Kaufmann Publishers Inc., USA, 432 pp.

Merchant S, Wood DE, Salzberg SL (2014) Unexpected cross-species contamination in genome sequencing projects. PeerJ 2: e675. https://doi.org/10.7717/peerj.675

Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA (2008) Database indexing for production MegaBLAST searches. Bioinformatics 24(16): 1757–1764. https://doi.org/10.1093/bioinformatics/btn322

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2018). Vegan: Community Ecology Package. R package version 2.5-4. https://CRAN.R-project.org/package=vegan

Pereira MB, Wallroth M, Jonsson V, Kristiansson E (2018) Comparison of normalization methods for the analysis of metagenomic gene abundance data. BMC Genomics 19: 274. https://doi.org/10.1186/s12864-018-4637-6

Piñol J, San Andrés V, Clare EL, Mir G, Symondson WOC (2014) A pragmatic approach to the analysis of diets of generalist predators: the use of next-generation sequencing with no blocking probes. Molecular Ecology Resources 14(1): 18–26. https://doi.org/10.1111/1755-0998.12156

Piñol J, Mir G, Gomez-Polo P, Agustí N (2015) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. Molecular Ecology Resources 15(4): 819–830. http://doi.org/10.1111/1755-0998.12355

Piñol J, Senar MA, Symondson WOC (2019) The choice of universal primers and the characteristics of the species mixture determines when DNA metabarcoding can be quantitative. Molecular Ecology 28(2): 407–419. https://doi.org/10.1111/mec.14776

Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P (2012) Who is eating what: Diet assessment using next generation sequencing. Molecular Ecology 21(8): 1931–1950. http://doi.org/10.1111/j.1365-294X.2011.05403.x

Pornon A, Escaravage N, Burrus M, Holota H, Khimoun A, Mariette J, Pellizzari C, Iribar A, Etienne R, Taberlet P, Vidal M, Winterton P, Zinger L, Andalo C (2016) Using metabarcoding to reveal and

quantify plant-pollinator interactions. Scientific Reports 6: 27282. https://doi.org/10.1038/srep27282

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, Goldsmith MR, Lawson D, Okamuro J, Robertson HM, Schneider DJ (2011) Creating a buzz about insect genomes. Science 331 (6023): 1386. http://doi.org/10.1126/science.331.6023.1386

RStudio Team (2015) RStudio: Integrated Development for R. RStudio, Inc, Boston, MA. http://www.rstudio.com/

Saitoh S, Aoyama H, Fujii S, Sunagawa H, Nagahama H, Akutsu M, Shinzato N, Kaneko N, Nakamori T (2016) A quantitative protocol for DNA metabarcoding of springtails (Collembola). Genome 59(9): 705–723. https://doi.org/10.1139/gen-2015-0228

Sarmashghi S, Bohmann K, Gilbert MTP, Bafna V, Mirarab S (2019). Skmer: Assembly-free and alignment-free sample identification using genome skims. Genome Biology 20: 34. http://doi.org/10.1186/s13059-019-1632-4

Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. Molecular Ecology Resources 15(6): 1289–1303. http://doi.org/10.1111/1755-0998.12402

Simhadri RK, Fast EM, Guo R, Schultz MJ, Vaisman N, Ortiz L, Bybee J, Slatko BE, Frydman HM (2017) The gut commensal microbiome of *Drosophila melanogaster* is modified by the endosymbiont *Wolbachia*. mSphere 2(5): e00287-17. https://doi.org/10.1128/mSphere.00287-17

Singh B, Crippen TL, Zheng L, Fields AT, Yu Z, Ma Q, Wood TK, Dowd SE, Flores M, Tomberlin JK, Tarone AM (2015) A metagenomic assessment of the bacteria associated with *Lucilia sericata* and *Lucilia cuprina* (Diptera: Calliphoridae). Applied Microbiology and Biotechnology 99(2): 869–883. https://doi.org/10.1007/s00253-014-6115-7

Srivathsan A, Sha JCM, Vogler AP, Meier R (2015) Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). Molecular Ecology Resources 15(2): 250–261. https://doi.org/10.1111/1755-0998.12302

Taberlet P, Bonin A, Zinger L, Coissac E (2018) Environmental DNA: For Biodiversity Research and Monitoring. Oxford University Press, UK, 268 pp. http://doi.org/10.1093/oso/9780198767220.001.0001

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21(8): 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C, Bruce C, Nevard T, Potts SG, Zhou X, Yu DW (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics. Methods in Ecology and Evolution 6: 1034–1043. https://doi.org/10.1111/2041-210X.12416

Tang M, Tan M, Meng G, Yang S, Su X, Liu S, Song W, Li Y, Wu Q, Zhang A, Zhou X (2014) Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics. Nucleic Acids Research 42(22): e166. https://doi.org/10.1093/nar/gku917

Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to analysis. Microbial Informatics and Experimentation 2: 3. https://doi.org/10.1186/2042-5783-2-3

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2007) Database resources of the National Center for Biotechnology Information, Nucleic Acids Research 35(1): D5–D12. https://doi.org/10.1093/nar/gkl1031

Wickham H (2016) ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York, 166 pp. http://ggplot2.org

Williams RS (2017) What's next? [The end of Moore's law]. Computing in Science & Engineering. 19(2): 7–13. https://doi.org/10.1109/MCSE.2017.31

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15: R46. https://doi.org/10.1186/gb-2014-15-3-r46

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3: 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zepeda Mendoza ML, Sicheritz-Pontén T, Gilbert MTP (2015) Environmental genes and genomes: understanding the differences and challenges in the approaches and software for their analyses. Briefings in Bioinformatics 16(5): 745–758. https://doi.org/10.1093/bib/bbv001

Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. Gigascience 2(1): 4. https://doi.org/10.1186/2047-217X-2-4

Supplementary material 1

Figure showing the comparison of the training and test datasets

Authors: Lidia Garrido-Sanz, Miquel Àngel Senar, Josep Piñol

Data type: Boxplot

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.2.48281.suppl1

Supplementary material 2

Additional information of the reference genomes

Authors: Lidia Garrido-Sanz, Miquel Àngel Senar, Josep Piñol

Data type: Excel table

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.2.48281.suppl2

Supplementary material 3
Relative species abundances for single-species libraries of the first run using different γ–δ combinations
Authors: Lidia Garrido-Sanz, Miquel Àngel Senar, Josep Piñol
Data type: Excel table
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.2.48281.suppl3

Supplementary material 4
Relative species abundances for single-species libraries of the second run using different γ–δ combinations
Authors: Lidia Garrido-Sanz, Miquel Àngel Senar, Josep Piñol
Data type: Excel table
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.2.48281.suppl4

Supplementary material 5
Relative species abundances for mixed-species libraries using the optimised γ–δ combination
Authors: Lidia Garrido-Sanz, Miquel Àngel Senar, Josep Piñol
Data type: Excel table
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.2.48281.suppl5

Supplementary material 6
Relative species abundances of the four single-species libraries sequenced twice
Authors: Lidia Garrido-Sanz, Miquel Àngel Senar, Josep Piñol
Data type: Excel table
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.2.48281.suppl6

Supplementary material 7
Assignment of the wrong identified species by "megablast" using nt NCBI database
Authors: Lidia Garrido-Sanz, Miquel Àngel Senar, Josep Piñol
Data type: Excel table
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.2.48281.suppl7

Supplementary material 8
Execution times of each tool in the complete pipeline for each mixed-species library
Authors: Lidia Garrido-Sanz, Miquel Àngel Senar, Josep Piñol
Data type: Excel table
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.2.48281.suppl8